Data Science: Profession and Education

Longbing Cao
University of Technology Sydney

Advanced analytics, data science, and new-generation artificial intelligence (AI) represent probably the most promising areas and directions in today's and near-future's Information and Communications Technology and Science, Engineering and Technology sectors and disciplines, where data science has become the major driving force of the new-generation AI. Data science and new-generation AI have attracted increasing interest from major governments, vendors, and academia, with important initiatives launched by major countries such as the United States,[1] China,[2] and the European Commission.[3]

However, despite the fact that the role of data scientists has been described as the sexiest job in the 21st century,[4,5] the qualifications and capabilities of a data scientist are not clearly defined; it is important yet undermined to define what makes the next-generation data scientists who can transform today's and future science, technology, innovation, and economy.[6]

On one hand, as an increasing number of data science courses are being created by online course offers and traditional institutions,[7] data science has clearly formed a new profession.[8,9,10] On the other, high-end data science employers (including major innovators, vendors, and enterprises) complain the limited availability of qualified data scientists to enable their strategic development and foster their future competitive advantages. These disclose the deviated status of existing professional and educational markets, the restricted benchmarking and accreditation of responsibilities and capabilities of data scientists, and the urgent need of standardizing and upgrading competencies and maturity of data science qualifications and education. This article intends to address these important issues, with an aim to contributing to the standardization and formalization of next-generation data science profession and education.

Digital Object Identifier 10.1109/MIS.2019.2936705
Date of current version 18 November 2019.

DATA SCIENCE AS A PROFESSION

Data science has driven the emergence of a new profession: *data science profession*, or simply *data profession*.[6] Typical evidence for this data profession formation includes the increasing number of clearly divided, diversified, well-defined and predominant data-oriented roles and responsibilities, the increasingly clearly articulated and pursued qualifications, the increasingly clarified competency and capabilities of these roles, and their increasingly profound impact on driving and transforming the data research, innovation, economy, and society.

Published by the IEEE Computer Society

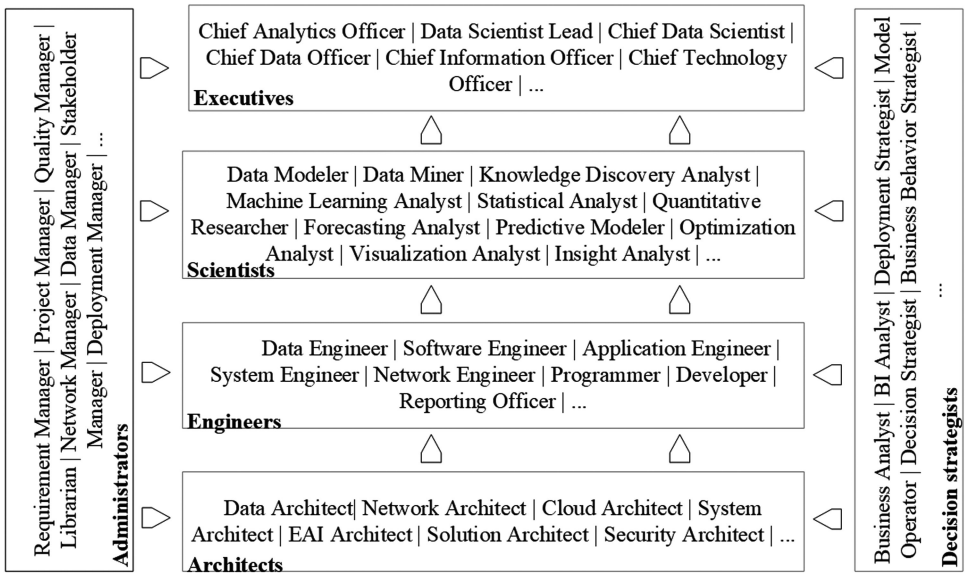

Figure 1. Data roles for enterprise data science.

Spectrum of Data Roles

Data roles broadly refer to job positions and responsibilities that are centered on, related to, or enabled by data and data-oriented (including data-driven, data-enabled, and data-based) tasks and agenda. Specifically, *data science roles* are those centered on data-driven discovery, innovation, and practices. They have been built on the profession of software engineering, database administration, business intelligence, computing, enterprise application integration, business analysis, and statistical analysis.

There are various perspectives to structure and categorize data roles. In conducting enterprise data science, an enterprise data science team is often formed with different roles for fulfilling scientific and engineering tasks. Typically, a data science team consists of data roles that fulfill the responsibilities and tasks for undertaking 1) the *infrastructure* for data computing, analytics, and decision support; 2) the *engineering* of data, software, applications, networking, communication, analytics, reporting, and decision support; 3) the *discovery* for data-driven scientific exploration, modeling, and optimization, including general and specific analytics, mining, learning, recognition, prediction, and refinement; 4) the *decision strategy* for designing, implementing, and evaluating data-driven decisions, strategies, and actions; 5) the *leadership* for planning, overseeing, and governing hierarchical (e.g., from enterprise-level to groups and teams in the enterprise) visions, missions, plans, and strategies; and 6) the *management* for administering data, resources, tasks, processes, risk, etc.

In Figure 1, various roles and positions are listed that address the abovementioned areas of responsibilities in a corporate data science group.[6] Typical data science roles that differ from existing BI professionals are data planning strategist, analytical architect, data modeler, subject-specific analyst (e.g., behavior analyst, financial analyst, and visual analyst), discovery scientist, model operator, quality assurer, decision strategist, etc. Interested readers are encouraged to read references such as the book by Cao[6] for more discussion on the various job positions, and in particular Chapter 10 "Data Profession" in the book by Cao[6] for discussion on the responsibilities associated with each of the abovementioned areas and those in Figure 1.

Data Scientists and Engineers

Among the various data roles, two major ones are data scientists and data engineers. They constitute a collaborative and functional data science team for conducting respective data science and data engineering tasks in an enterprise.

There have been intensive discussion on roles and responsibilities of data scientists,[6,9] which are usually either broad-based or specific.

So, who are data scientists and data engineers, respectively? *Data scientists* are those data professionals whose major responsibilities and agenda are centered on data-driven exploration and discovery. *Data engineers* instead focus on preparing and processing data and related supporting facilities for data-driven discovery and for fulfilling data-driven discovery results. In short, data scientists make sense of data for discovering data values and insights, whereas data engineers get data ready, support the discovery, and implement the methods and tools for discovering data value and insight. Figure 2 summarizes the responsibilities and qualifications of data scientists and data engineers, respectively.

What do data scientists and data engineers do, respectively? The main responsibilities of and tasks conducted by data scientists consist of the following: 1) understanding problem complexities; 2) identifying and specifying constraints and requirements; 3) understanding and quantifying data characteristics; 4) translating underlying challenges to analytical problems; 5) planing analytical strategies and design; 6) conducting data exploration and discovery; 7) evaluating and optimizing analytical results; 8) extracting data value and insight; 9) communicating and interpreting results with stakeholders; and 10) operationalizing data exploration and discovery.

In contrast, data engineers are responsible for 1) understanding the business domain and problems; 2) conducting business requirement analysis; 3) getting data ready; 4) getting data system ready; 5) ensuring data quality and ethical compliance; 6) programming data models; 7) computing data; 8) manipulating analytical results; 9) managing projects; and 10) testing data system and assuring the system quality to achieve design objectives. In practice, much of the abovementioned work may be shared and collaboratively undertaken between data scientists and engineers.

What qualifications should data scientists and data engineers have to fulfill the aforementioned respective responsibilities and tasks? The following qualifications and capabilities may be required for data scientists: 1) data science thinking; 2) scientific leadership; 3) doctoral qualifications in related fields; 4) capabilities of handling complex systems and problem solving; 5) solid foundation in statistics, analytics, and learning; 6) hands-on experience in computing; 7) good collaborative, organizational, and communication skills; and 8) rich interdisciplinary knowledge and cross-domain experience.

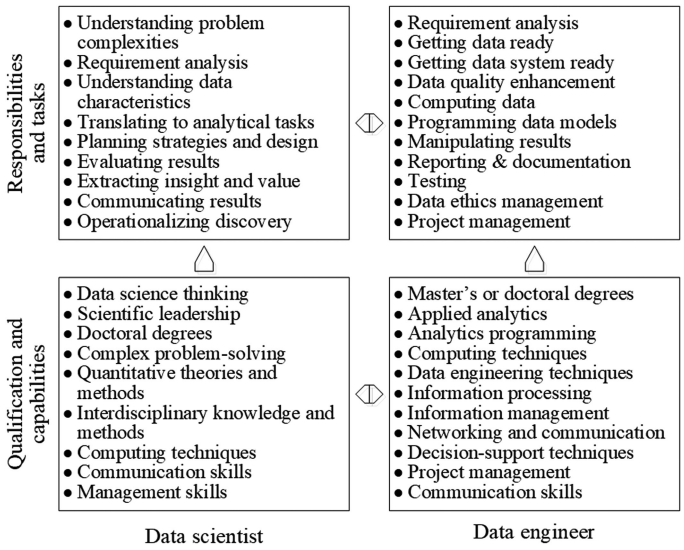

Figure 2. Responsibilities and qualifications of data scientists and engineers.

Data engineers may hold the following qualifications and skills: 1) methodologies of computing and engineering; 2) master's or doctoral qualifications in related fields; 3) knowledge of applied statistics; 4) knowledge and skills in software engineering, information systems, programming, distributed and cloud computing, high-performance computing, networking and communication, information retrieval, information security, and enterprise application integration, etc.; 5) skills and practice in data processing; 6) experience of system design, implementation, testing, and deployment; 7) knowledge and experience of project management; 8) skill and experience of user interface design and decision-support systems; and 9) good collaborative, organizational, and communication skills.

DATA SCIENCE COMPETENCIES

The competencies for enabling data science as a profession and conducting creative data

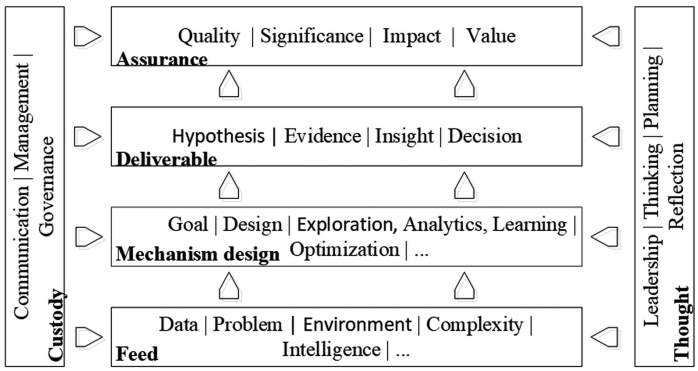

Figure 3. Data science view and structure.

science research and actionable practice require us to establish data science thinking and data science knowledge base and skill set. Their maturity in an individual or an organization determines how better the individual or organization is than others, and how farther and deeper their data science can bring about to them.

Data Science Thinking

What makes data science a new science? In many possible technical answers to and discussion on this important question, the utmost is the data science thinking.[6] The critical foundation and success factor of data science is "what and how to think about data"; here, the thinking goes beyond creative and critical thinking and data analytical thinking. *Data science thinking* refers to the cognitive and methodological perspectives, thinking traits and habits, and design paradigms and strategies of the mind in handling data problems and systems.

First, from the cognitive and methodological perspectives, data science is a higher level scientific field, a transdisciplinary science, a complex system, and a comprehensive cognitive and discovery process (see more specific discussion on these arguments in the book by Cao[6]). Accordingly, a systematic view of the scientific methodologies, disciplinary structure, problem complexities, and problem-solving thinking, and technical approaches is essential. The integrative systematism[11] may be required to synthesize bottom-up reductionism and top-down holism to generate a systematic view and enable a systematological process of data science. Transdisciplinary data science synthesizes and transforms multiple relevant sciences and fields such as informatics, computing, statistics, and sociology to a new field.

Accordingly, a comprehensive view and structure of data science can be created, such as that shown in Figure 3.[6] The *data science thinking* views a data problem and system as composed of four interconnected and progressive layers: 1) the feed layer, 2) the mechanism design layer, 3) the deliverable layer, and 4) the assurance layer; and two enablers: 1) the thought wing and 2) the custody wing.[6] This comprehensive understanding of data science starts from the identification, quantification, and formalization of system complexities of data problems and systems, which involves X-complexities and X-intelligence w.r.t. data, domain, organization, society, human, network, and behavior.[6,12]

Second, the abovementioned cognitive and methodological understanding of data science has to be enabled by general and specific data-oriented thinking traits and habits. The exploration of these X-complexities and X-intelligence and their metasynthesis[11] requires the establishment of scientific thinking for data science, which is complexity- and intelligence-driven, data-centric, exploratory and analytical, evidence- and fact-based, and reproducible and interpretable. Many critical, creative, and systematic thinking traits and skills are required to achieve this data science thinking, e.g., the inquisitiveness and imaginative thinking about the *unknown world*[12] w.r.t. unknown data challenges, problems, complexities, hierarchy, structures, distributions, relations, heterogeneities, and unknown capabilities, opportunities, and solutions; and the integrity, soundness and interpretability of data-driven exploration processes, analytical and learning theories and methods, and resultant evidence.

Lastly, data-centric design paradigms, strategies, and patterns are necessary to produce original, interesting, and actionable[13] data products.[6] It has been one of the most important agendas in the relevant data science communities to invent appropriate design paradigms, strategies, patterns, architectures, and models to address common or specific data characteristics, system complexities, and outcome expectations. This has been reflected in the journey of statistical analysis, machine learning, knowledge discovery,

and broad informatics and analytics, as exemplified by classic design strategies and patterns such as complementation versus contrast, individual versus hybridization design, coupling versus disentanglement, breadth versus depth, and structuring (e.g., partition, graph, tree, and ensemble committee); and more recent effort on deepening network depth for creating deep neural networks and incorporating design strategies such as attention, memory, gating, and adversary into deep networks. Such effort continues on creating new and more effective and efficient design patterns and strategies to tackle more sophisticated data characteristics and complexities.

Data Science Knowledge and Skills

The different data roles including data scientists and data engineers are empowered by respective bodies of knowledge and set of skills and capabilities. Since data roles are comprehensive (spreading over the whole spectrum of making sense of data and delivering data value) and specific (each role is responsible for particular tasks and expectations), the relevant knowledge map and skill set are broad and can be categorized in terms of different criteria and purposes. Figure 4 illustrates a framework of competencies for data science teams and individual data scientists.[6] The competency framework suggests data science education and training to foster a comprehensive knowledge and capability set for qualified all-rounded data scientists, including key knowledge, skills, and experience in data science thinking, foundational theories, engineering techniques, work-ready practice, communication skills, management, and leadership.

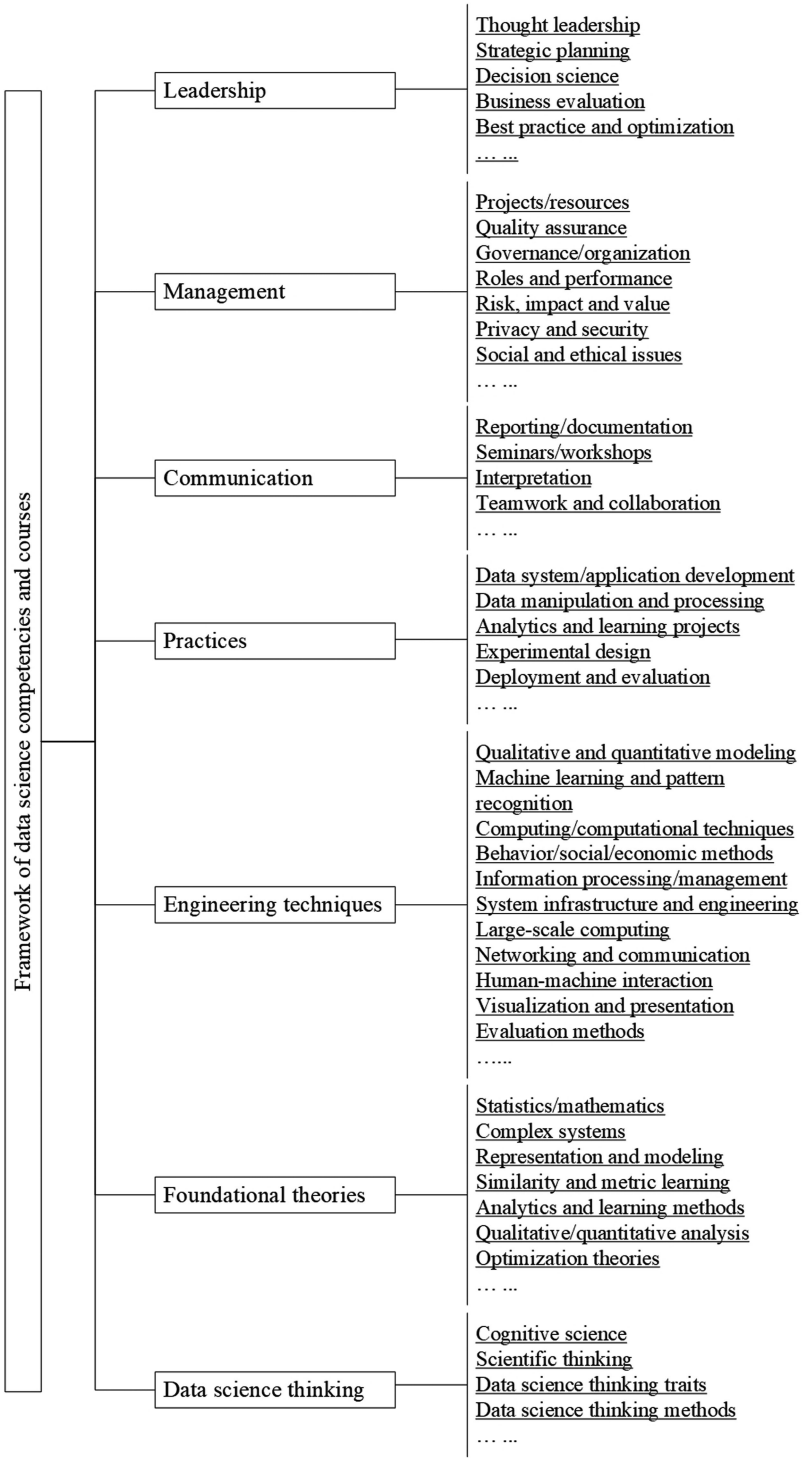

Figure 4. Framework of data science competencies and courses.

First, it is important to train all data roles with *data science thinking* and mental capabilities for them to think with data.[14] The required methodologies, research methods, cognitive skills, and mental traits for 1) enabling general

cognitive and scientific thinking, including creative and critical thinking, induction and reduction, abstraction and summarization, and logical and imaginative thinking; and 2) developing specific data science thinking traits and methods, including statistical thinking, computational thinking, data analytical thinking, learning and inferential thinking, and optimization and ethics.

Second, the *foundational theories* for data science may involve different disciplines and areas, typically including 1) the theories in statistics, mathematics, and complex systems for understanding problem complexities and data characteristics; 2) the theories and methods for representation, exploration, analysis, learning, and modeling of complex data, behaviors, and problems; and 3) the theories and methods for quantifying similarity, dissimilarity, quality, impact, and risk; and 4) the theories for evaluating and optimizing learnability, and computational complexity.

Furthermore, many *engineering techniques* have to be involved in data science and engineering. These include the following: 1) qualitative and quantitative analytical and learning techniques, typically including statistical analysis, knowledge discovery, machine intelligence learning, pattern recognition, natural language processing, and multimedia data analysis, to build and optimize analytical and computational algorithms and models for descriptive, diagnostic, and predictive and prescriptive analytics and learning; 2) information processing and management techniques and tools for data preparation, exploratory analysis, information retrieval, data engineering, and analytics programming; and 3) system design, infrastructure building, and decision-supporting techniques and tools for enabling human–machine interaction, data infrastructure and platform construction, large-scale computing (including high-performance analytics, distributed analytics, and cloud analytics), networking and communication, visualization, and presentation.

In addition, *work-ready practices* are critical for converting data and data science to value and productivity by applying the abovementioned theories and techniques. Better practices are built on practical skills and hands-on experience of undertaking impactful enterprise data science innovation, actionable and interpretable experimental designs, effective and reproducible analytics projects, exemplary and transferable case studies, and successful widespread applications.

Lastly, the success of data science depends not only on original theories, innovative techniques, and actionable practices, but also on effective communications, management, and leadership. Communication skills are critical for building a high-performing data science team and delivering data-science-driven strategic values and decision-making actions. *Data science communications* may involve regular workshops and seminars; professional reporting, documentation, and visualization of data science design, modeling, processes, and results; informative storytelling and evaluation, and business-friendly interpretation of data-driven discovery findings; and team-based collaboration and reflection.

Data science management involves techniques and capabilities for conducting effective data organization, governance, and auditing; management of data, resources, projects, roles and responsibilities, and results; assurance of mitigating privacy, security, risk, impact, and social and ethical issues; and optimization of operations, deployment, and decisions. In contrast, *data science leadership* roles hold the knowledge and experience of enterprise data science thinking, research and innovation leadership, strategic planning, decision science, business evaluation, risk management, best practice, optimization, etc.

Data Science Competency Maturity

In reality, the data science competencies between individuals and between organizations are usually divided. The competency imbalance reflects the difference and gaps between organizations which undertake data science.

The *data science competency maturity* refers to the level of an individual's or an organization's capability and capacity for undertaking best possible data science research and best practice. A data science competency maturity model builds benchmarks to structure and quantify the maturity level of competency, which can be measured in terms of data maturity, data science capability maturity.[6] Figure 5 summarizes

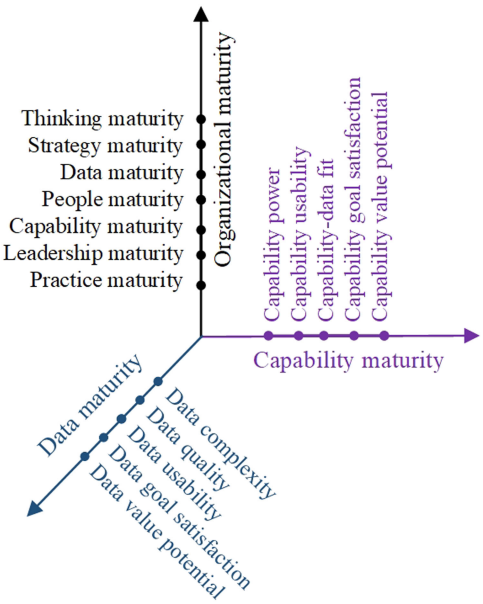

Figure 5. Data science competency maturity.

various aspects of competency maturity in data science.

Data maturity measures the level of data complexities, the quality of data, the usability of data, the satisfaction rate of data for achieving anticipated objectives, and the potential of a data value. These aspects are specified and quantified in terms of target sources of data, which may incur differences, e.g., in some corporate organizations, partial data may be highly mature while the global set may not be.

Data science capability maturity refers to the level of knowledge, capability, practice, and experience of either individual data roles or a data science team for manipulating data and delivering best possible outcomes. The capability maturity can be structured and quantified in terms of the sufficiency and power levels of essential capabilities, the usability of capabilities, the fit level between the capabilities held and those required for utilizing data maturity, achieving data science objectives, and maximizing the value potential of data and capabilities.

Lastly, *data science organizational maturity* refers to the level of organizational maturity in undertaking data science research, innovation, and practice. This is further depicted and categorized in terms of the maturity of strategic organizational data science thinking (thinking maturity), the maturity of data science strategic planning and policies (strategy maturity), organizational infrastructures and capabilities for fulfilling data science (capability maturity), the maturity of organizational data (data maturity), the maturity of organizational data science team (people maturity), the level of practical experience (practice maturity), and the excellence level of data science organizational leadership (leadership maturity), etc. In addition, data science organizational maturity can be further categorized in terms of organizational hierarchy, e.g., the corporate, departments or business units, teams, and individuals.

DATA SCIENCE EDUCATION

There are a tremendous number of courses available online and in academic and training institutions to teach basic to advanced subjects. However, their quality is highly divided, and few of them have formed the systematic and intrinsic strategies and programs to train the next-generation data scientists and to fit the strategic need of booming data economy and data science profession. Here, we briefly review the gaps, and introduce a framework for awarded data science qualifications and pathways to train qualified data professionals.

Gaps in Data Science Education

In recent years, hundreds of courses and subjects have been offered at academic institutions and training organizations.[9] These courses are offered in core disciplines of statistics and mathematics, computer science, business and management, and side ones such as disciplines of environmental science and finance. It has been observed that almost every major university either has opened or is on the way to open courses in data science. In addition, online courses, training courses, and open courses are increasingly diversified and attracting teaching resource shift from academic institutions to their course-offering channels.

The body of knowledge delivered in typical on-campus data science courses include major subjects such as (applied) statistics, AI, machine learning, pattern recognition, big data analytics, and data mining; and minor ones such as programming, visualization, business analytics,

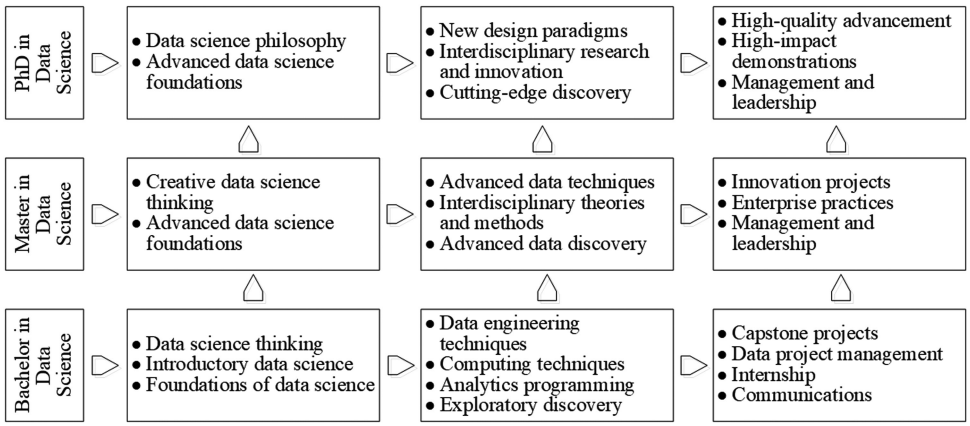

Figure 6. A framework for data science qualifications and pathways.

project management, and capstone projects. These subjects may be offered to both undergraduate and postgraduate students.

A comprehensive review of the relevant courses and subjects offered for data science comes up with the following observations and gaps[6]: 1) none to few courses teach students how to think about data, i.e., the skills and capabilities of data science thinking; 2) truly transdisciplinary data science courses are few, most of the courses form a "miscellaneous decoupled platter" of existing subjects delivered by different faculties; 3) the level of course quality is very divided, and the same for the level of knowledge advancement and the coverage and sufficiency of data science competency; 4) the independent data-driven problem-solving research and innovation capabilities are missing yet critical for conducting real-life data science; and 5) there are serious gaps between the challenges of real-world problem complexities and the advancement of knowledge taught in most of the courses available.

Data Science Qualification Framework

Some major questions to be asked in constructing a pragmatic data science course framework include the following: 1) What makes a qualified next-generation data scientist? (2) What are the gaps and issues in existing data science courses? (3) What are available to foster our next-generation data scientists? (4) How to enable students to cope with unknown challenges and unavailable knowledge and invent new knowledge required?

It has to be a joint effort of relevant government, industry, academic institutions, and individuals to systematically address the abovementioned challenges. Various respective resources and efforts in academia, industry (including corporate training and public courses), and policy-making bodies have to be committed. Here, we want to focus on how to transform and structure the awarded courses of data science since they play the driving role in training scientists and engineers. In Figure 6, a framework of data science qualifications is presented, which paves paths from undergraduate to master's and doctoral studies.

A *Bachelor in Data Science* may have built knowledge, capabilities, and experience through three-stage training. Stage 1 is to learn data science fundamentals, which foster students with data science thinking, introductory data science, and foundations of data science. Stage 2 transfers knowledge and engineering skills to students for undertaking data processing, engineering, computing, and programming, as well as exploratory discovery. Stage 3 focuses on data science practices, which involve multiple capstone projects, internships, and applications of data science; this will enable and enhance the skills and experience for assuring sound data understanding results, introducing project management, managing social and ethical issues, and communicating actionable results to stakeholders. As a result, graduates are ready for workplace in enterprise data science as data analysts or business analysts.

Furthermore, a *Master in Data Science* is trained to be a data specialist in a specific

business domain or a research area of data science through three-stage training. In Stage 1, students are built with advanced foundations in terms of creative data science thinking and foundations for advanced data science. Stage 2 transfers the advances in data science research and development to students for them to grasp the latest and most advanced techniques, and to conduct advanced data discovery; students are also empowered with the relevant interdisciplinary theories and methods. Stage 3 fosters students with advanced data science innovation capabilities, enterprise data science innovation and practices, and knowledge and skills of management and leadership.

Lastly, a *PhD in Data Science* is trained to be a senior data science leader, who can create profound and rigorous theoretical breakthroughs, and has in-depth and original understanding and processing of sophisticated and previously unknown data and problem complexities. Accordingly, doctoral students focus on developing original and creative philosophical thinking in data science, transforming and creating scientific paradigms for data science research and innovation, and performing novel and significant interdisciplinary research and innovation. They are expected to produce unique and significant research leadership and knowledge advancements, and demonstrating better practice of translating high-quality data science advancements into high-impact demonstrations.

CRITICAL AGENDA

Data science has been a driving new productivity of enabling and creating new science, development, applications, and economy.[6] Accordingly, data science profession has emerged as a new profession in which data scientists play the profound roles, and data science education should train the next-generation data scientists to achieve the aforementioned agenda.

The systematic and quality development of data science profession and education relies on setting up and executing significant professional and national science and innovation agenda, for example:

- Comprehensive market surveys of the market demand and supply, the quality and satisfaction of existing generation of data roles, and the gaps between them and the fast-developing market need especially that of major data-based innovators, vendors, and enterprise end users.
- Professional benchmarking and accreditation of data science qualifications and courses that conform to the joint agreement with the relevant disciplines, industry, professional bodies, and policy-making institutions.
- Regional and global collaborations in developing standard guidance and curricula outlines that minimize the deviation and maximize the positive experience between institutions, countries, and regions.
- Educational and training initiatives and programs that support transcurricula programs to train data scientists with compound transdisciplinary knowledge and cross-domain experience.
- New educational and training strategies, plans, and programs that substantially enhance the qualifications and capabilities of the existing professionals and train high-calibre next-generation data scientists.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council Discovery Grant (DP190101079).

REFERENCES

1. D. J. Trump, "Executive order on maintaining American leadership in artificial intelligence," Feb. 2019. [Online]. Available: https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/
2. The State Council, The People's Republication of China, "Notice of the state council on issuing the development plan on the new generation of artificial intelligence," (in Chinese), Jul. 2017. [Online]. Available: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm
3. European Commission, "Artificial intelligence for europe," Apr. 2018. [Online]. Available: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625
4. T. H. Davenport and D. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 70–76, 2012.

5. L. Cao, "Data science: Nature and pitfalls," *IEEE Intell. Syst.*, vol. 31, no. 5, pp. 66–75, Sep./Oct. 2016.
6. L. Cao, *Data Science Thinking: The Next Scientific, Technological and Economic Revolution* (Data Analytics). New York, NY, USA: Springer Int. Publishing, 2018.
7. R. D. D. Veaux *et al.*, "Curriculum guidelines for undergraduate programs in data science," *Annu. Rev. Statist. Appl.*, vol. 4, no. 2, pp. 1–16, 2017.
8. M. A. Walker, "The professionalisation of data science," *Int. J. Data Sci.*, vol. 1, no. 1, pp. 7–16, 2015.
9. L. Cao, "Data science: A comprehensive overview," *ACM Comput. Survey*, vol. 50, 2017, Art. no. 43.
10. UK Government, "Digital, data and technology profession capability framework," Dec. 2018. [Online]. Available: https://www.gov.uk/government/collections/digital-data-and-technology-p rofession-capability-framework
11. L. Cao, *Metasynthetic Computing and Engineering of Complex Systems*. New York, NY, USA: Springer, 2015.
12. L. Cao, "Data science: challenges and directions," *Commun. ACM*, vol. 60, no. 8, pp. 59–68, 2017.
13. L. Cao, P. S. Yu, C. Zhang, and Y. Zhao, *Domain Driven Data Mining*. New York, NY, USA: Springer, 2010.
14. B. Baumer, "A data science course for undergraduates: Thinking with data," *Amer. Statistician*, vol. 69, no. 4, pp. 334–342, 2015.